# DualAvatar: Robust Gaussian Avatar with Dual Representation


Jinsong Zhang
jinszhang@tju.edu.cn
Tianjin University
Tianjin, China

I-Chao Shen
jdilyshen@gmail.com
The University of Tokyo
Tokyo, Japan

Jotaro Sakamiya
jotarosakamiya@g.ecc.u-tokyo.ac.jp
The University of Tokyo
Tokyo, Japan

Yu-Kun Lai
yukun.lai@cs.cardiff.ac.uk
Cardiff University
Cardiff, United Kingdom

Takeo Igarashi*
takeo@acm.org
The University of Tokyo
Tokyo, Japan

Kun Li*
lik@tju.edu.cn
Tianjin University
Tianjin, China


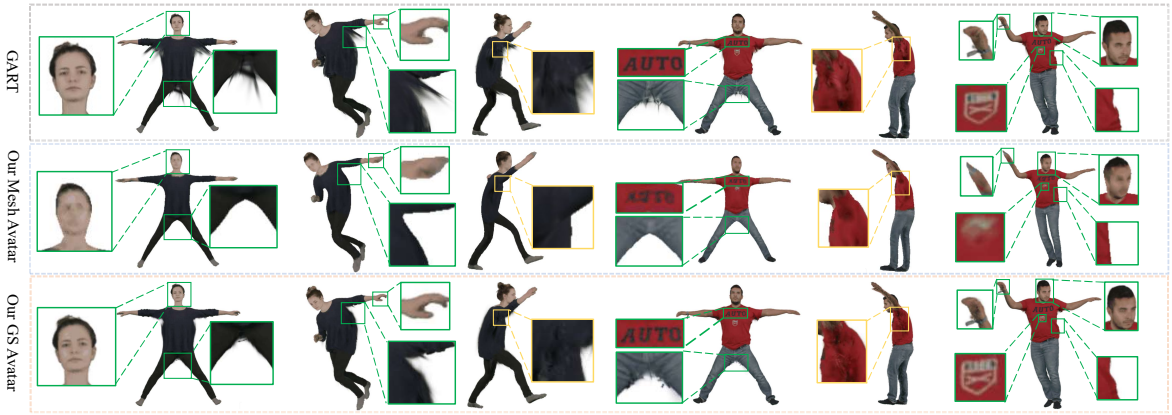


**Figure 1: Avatar reconstruction and animation results using 20 images. Compared with GART [Lei et al. 2024], for each subject, the first column shows the GS avatar in the canonical space, and the other two columns show the animation results.**

## 1 INTRODUCTION

Creating 3D human avatars from monocular videos has significant potential value in virtual reality, gaming, and movie production. However, creating a robust 3D animatable avatar for novel poses with limited observations remains challenging. Recent works utilizing 3D Gaussian Splatting (3DGS) [Kerbl et al. 2023] have made significant progress in reconstructing high-quality avatars from a single video. Among them, GART (Gaussian Articulated Template Model) [Lei et al. 2024] models deformable geometry and appearance using dynamic 3D Gaussians explicitly. They achieve state-of-the-art performance in monocular human reconstruction and novel view synthesis with real-time rendering. However, their method encounters difficulties in reconstructing a robust avatar for novel poses due to limited human poses observed in the training monocular video. As a result, their method generates unsatisfactory rendering results with artifacts (first row in Figure 1).

Another line of work adopts the mesh representation to reconstruct robust mesh avatars for novel poses. However, due to its limited resolution and fixed topology, the mesh avatar often has inaccurate geometry and blurry texture (second row in Figure 1).

To address these problems, we propose *DualAvatar*, a method for reconstructing a robust Gaussian Splatting (GS) avatar for novel poses from a single monocular video. Our method is based on the observation that while the mesh-based avatars may have low-quality appearance, they are capable of adapting to novel poses effectively. Therefore, during training, we propose to optimize two avatars concurrently using two different representations: a GS avatar and a mesh avatar. This allows us to refine the unseen regions and poses of the GS avatar, such as the armpits, with guidance from the mesh avatar. As a result, we can generate a robust GS avatar that can adapt to new poses (third row in Figure 1).

## 2 METHOD

Given a monocular video of a person rotating in an A-pose, our goal is to reconstruct a robust GS avatar that has fewer artifacts under unseen poses. Figure 2 shows the overview of our method. The


*Corresponding author

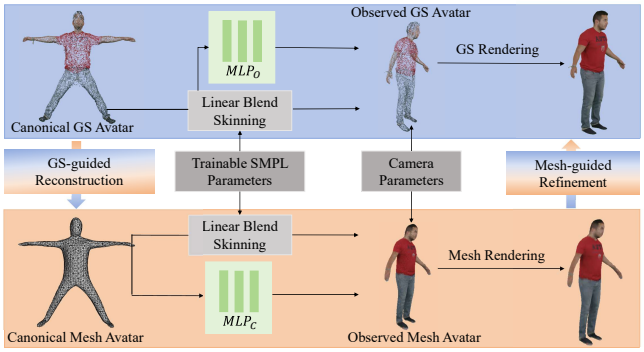


**Figure 2: The overview of our method.**

mesh avatar and the GS avatar are optimized in canonical space. With predicted pose parameters, we deform the avatars and render the outputs. By minimizing the distance between the rendered image and the target image, our DualAvatar can be optimized end-to-end. During this process, we propose GS-guided reconstruction and mesh-guided refinement to obtain a robust GS avatar.

**Avatar Representation.** We adopt 2D GS [Huang et al. 2024] as our avatar representation and represent each Gaussian $g_i = \mu_i, r_i, s_i, c_i, o_i$ using its position, rotation, scale, color and opacity. The mesh avatar is parameterized using DMTet [Shen et al. 2021]. Specifically, the mesh avatar is derived from a tetrahedral mesh $TM$ with $N$ vertices and $M$ tetrahedra. By employing a learnable signed distance field (SDF), we can compute the SDF value for each vertex in $TM$ to obtain a triangle mesh avatar $MA = (\mathbf{V}, \mathbf{F})$, where $\mathbf{V}$ and $\mathbf{F}$ denote the positions of the vertices and the faces. To render color images, we predict the color, *i.e.*, RGB value, $\mathbf{V_c} = \mathrm{MLP}_C(\mathbf{V})$ of each vertex using a hash table and a linear layer.

**Avatar Optimization.** Given the trainable pose parameters $\theta$, we need to deform the canonical avatars to the target pose and render the images. For the mesh avatar $MA$, we treat each vertex as a point and apply linear blend skinning to deform the avatar. Specifically, for each vertex $v_i$ in $\mathbf{V}$, the deformed vertex $v_i^o$ in the observation space is defined as:

$$v_i^o = Tv_i, T = LBS(v_i, \theta) = \sum_{k=1}^{K} W_k(v_i) B_k(\theta), \tag{1}$$

where $T$ represents the transformation matrix, $K$ denotes the number of rigid bones, $W_k$ and $B_k$ are the learnable skinning weights and bone transformation based on $\theta$ of the $k$-th bone, respectively. To deform the GS avatar, we first apply linear blend skinning as the mesh avatar, *i.e.*, $\mu_i^{'} = T\mu_i$ and $r_i^o = \tilde{T} r_i$, where $\tilde{T}$ represents the rotation matrix in $T$. Then, we adopt two linear layers to predict the deformed positions of Gaussians: $\mu_i^o = \mu_i^{'} + MLP_o(\mu_i^{'})$.

We then render the GS avatar using Gaussian rasterization [Huang et al. 2024] and the mesh avatar using Nvdiffrast [Laine et al. 2020]. We adopt $\mathcal{L}_1$ term and the D-SSIM term [Lei et al. 2024] between the rendered images of two avatars and the target images.

*GS-guided Reconstruction for Mesh Avatar.* To provide geometric information for the mesh avatar, during training, we leverage the GS avatar to constrain the geometry of the mesh avatar in the canonical space. Specifically, we sample 50,000 points $V_s$ from mesh avatar, and minimize the chamfer distance between $V_s$ and the positions $\mu$ of all Gaussians in GS avatar.

*Mesh-guided Refinement for GS Avatar.* To reduce artifacts in invisible regions in the reconstructed GS avatar, we adopt a simple but effective way to refine the GS avatar using the mesh avatar. We use a $\mathcal{L}_1$ regularization to make all pixels of the GS-rendered images similar to those of the mesh-rendered images in the canonical space, which can inpaint the invisible regions according to the mesh avatar.

**Table 1: Quantitative comparison on SnapshotPeople dataset.**

| | male-3-casual | | | | female-3-casual | | | |
|---|---|---|---|---|---|---|---|---|
| | Gaussians↓ | PSNR↑ | SSIM↑ | LPIPS↓ | Gaussians↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| GART (10 images) | 46.1 K | 28.48 | 0.9680 | 0.0345 | 36.8 K | 24.15 | 0.9536 | 0.0497 |
| **Ours** (10 images) | 24.0 K | 28.67 | 0.9674 | 0.0346 | 19.5 K | 24.83 | 0.9569 | 0.0481 |
| GART (20 images) | 46.4 K | 30.21 | 0.9766 | 0.0334 | 38.8 K | 25.25 | 0.9609 | 0.0454 |
| **Ours** (20 images) | 24.4 K | 30.03 | 0.9748 | 0.0319 | 21.0 K | 25.50 | 0.9614 | 0.0432 |
| | male-4-casual | | | | female-4-casual | | | |
| | Gaussians↓ | PSNR↑ | SSIM↑ | LPIPS↓ | Gaussians↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| GART (10 images) | 39.5 K | 26.53 | 0.9581 | 0.0535 | 44.1 K | 27.83 | 0.9643 | 0.0327 |
| **Ours** (10 images) | 22.6 K | 26.60 | 0.9577 | 0.0506 | 22.8 K | 28.06 | 0.9650 | 0.0329 |
| GART (20 images) | 41.6 K | 27.00 | 0.9639 | 0.0552 | 44.1 K | 28.76 | 0.9708 | 0.0323 |
| **Ours** (20 images) | 23.2 K | 27.04 | 0.9626 | 0.0514 | 23.7 K | 28.84 | 0.9704 | 0.0319 |

## 3 RESULTS

To evaluate the performance of DualAvatar, we conduct experiments on PeopleSnapshot dataset [Alldieck et al. 2018], which contains various monocular videos of a person rotating in an "A" pose. We use different numbers of images as training data to show the experimental results compared to GART [Lei et al. 2024]. As shown in Table 1, our method achieves superior performance across most metrics with half the Gaussian number. As shown in Figure 1 our model not only generates realistic images but also effectively render unseen pose results using only 20 images as training data. Moreover, our GS avatar can inpaint invisible regions under the guidance of the mesh avatar (highlighted in the yellow boxes). For dynamic results, please refer to the supplementary video.

## ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (62122058 and 62171317), the Science Fund for Distinguished Young Scholars of Tianjin (No. 22JCJQJC00040), and JSPS Grant-in-Aid JP23K16921, Japan.

## REFERENCES

Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018. Video based reconstruction of 3d people models. In *Proc. CVPR*. 8387–8397.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2024. 2D gaussian splatting for geometrically accurate radiance fields. *arXiv preprint arXiv:2403.17888* (2024).

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* 42, 4 (July 2023).

Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. 2020. Modular Primitives for High-Performance Differentiable Rendering. *ACM Trans. Graph.* 39, 6 (2020).

Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. 2024. GART: Gaussian articulated template models. In *Proc. CVPR*. 19876–19887.

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. 2021. Deep marching tetrahedra: a hybrid representation for high-resolution 3D shape synthesis. *Proc. NeurIPS* 34 (2021), 6087–6101.